效应量置信区间的原理及其实现[1]

王珺[1] 宋琼雅[1] 许岳培[2] 贾彬彬[3] 胡传鹏[4,5]

（[1]中山大学心理学系，广州，510006；
[2]上海师范大学教育学院，上海，200234；
[3]上海体育学院，上海，200438；
[4] Neuroimaging Center, Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Centre Mainz, 55131 Mainz, Germany；
[5] Deutsches Resilienz Zentrum (DRZ), University Medical Centre of the Johannes Gutenberg University, 55131 Mainz, Germany）

摘　要　在心理学可重复危机的背景之下，报告效应量及其置信区间正逐渐成为主流心理学界所要求的新标准，但是研究者可能对效应量的置信区间缺乏足够的理解。为增强研究者对效应量置信区间的理解及应用，本文介绍了心理学研究中最常用的效应量指标——Cohen's *d* 与 η^2——的置信区间的基本原理，即，在备择假设（H_1）为真时，需要通过迭代估计的方式来估计相应非中心分布的非中心分布参数，从而构建 Cohen's *d* 与 η^2 的置信区间。其中 Cohen's *d* 对应的是非中心 *t* 分布；而 η^2 对应的则是非中心 *F* 分布。使用现有的计算机程序，能够对 Cohen's *d* 与 η^2 的置信区间进行计算，例如 R 与 JASP，本文对此进行了分别展示。报告效应量置信区间不仅有助于研究者更好地进行统计推断，也有利于整个科学界知识的积累，因此本文介绍的方法对研究者具有十分重要的意义。

关键词　效应量；置信区间；Cohen's *d*；Eta squared；R

1　引言

统计推断是研究者根据数据进行逻辑推导从而验证研究假设的必要手段。虚无假设显著性检验(null hypothesis significance test, NHST)是心理学研究中最为常用的统计推断手段(Cumming et al., 2007)。但该方法以 *p* 值是否小于 0.05 作为决定统计显著性的指标，间接导致了心理学研究的假阳性过高；且 *p* 值受抽样样本的影响较大，不适合作为重复研究或跨实验研究比较的统计指标(胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平, 2016)。近年来，随着对心理学研究可重复性的广泛关注，NHST 的局限性再次引起众多学者的重视(Kline, 2004; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011)。为了弥补 NHST 的不足，新的统计方法开始逐渐被引入心理学研究，例如基于估计的统计(estimates-based statistics)(Cumming, 2012, 2014)、贝叶斯因子(胡传鹏, 孔祥祯, Wagenmakers, Ly, 彭凯平, 2018; Wagenmakers et al., 2018)、似然性方法(Etz, 2018)。其中，基于估计的统计方法由于易于理解，且能够弥补

[1] 通讯作者：胡传鹏, Email: hcp4715@hotmail.com。文中演示数据和代码可在线获取：https://osf.io/4ameb/。

NHST 的不足，被国内外研究者推荐。该方法所强调的效应量(effect size)及其置信区间(confidence intervals, CIs)正逐渐成为国际、国内重要心理学期刊论文中必须报告的统计指标(APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, 2018; Cumming, 2014)。

尽管如此，相比“统治”了心理学数十年的 NHST，效应量及其置信区间在心理学研究中的使用仍十分有限，极少研究报告效应量的置信区间(Fritz, Morris, & Richler, 2012)。国内研究者虽对效应量的概念进行过不少的介绍(胡竹菁, 2010; 卢谢峰, 唐源鸿, 曾凡梅, 2011;郑昊敏, 温忠麟, 吴艳, 2011)，但却很少提及效应量的置信区间。

值得注意的是，心理学专业研究人员、学生对置信区间仍有一定误解(胡传鹏等, 2016; Hoekstra, Morey, Rouder, & Wagenmakers, 2014)。例如，胡传鹏等人 (2016)针对国内研究者对 CI 的理解情况进行了调查。在该调查中，呈现一个假想的研究，其效应的 95%置信区间为[0.1, 0.4]，受访者需要判断是否能够根据这个置信区间推断出如下 6 个陈述：A，真实的均值大于 0 的可能性至少是 95%；B，真实的均值等于 0 的可能性小于 5%；C，真实的均值等于 0 的“零假设”很可能是不正确的；D，真实的均值有 95%的可能性在 0.1 和 0.4 之间；E，我们有 95%的信心认为真实的均值在 0.1 和 0.4 之间；F，如果我们重复该实验，则 95%的时候，真实的均值会在 0.1 和 0.4 之间。上述 6 个陈述均属于对置信区间的误解(Hoekstra et al., 2014)，但是大部分受访者或多或少将其判断为正确解读。（见图 1，数据来自 Lyu, Peng, & Hu, 2018）。实际上，置信区间的正确解读应该是，如果不断重复该实验并计算置信区间，在所有计算出来的置信区间中，约有 95%的置信区间包含真实的均值。因此这里的[0.1, 0.4]是理论上众多置信区间中的一个，其是否包括真值是未知的(Cumming, 2014)。

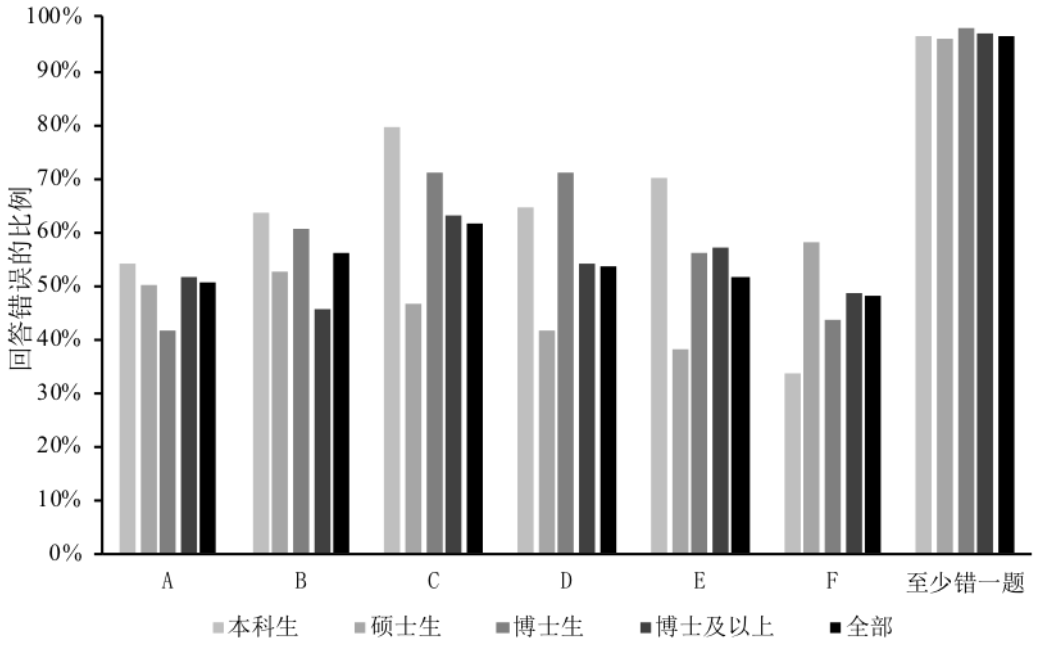

图 1 国内心理学学生及研究者在 6 个关于 CI 陈述上的错误率

为加深研究者对效应量及其置信区间的理解，同时便于研究者准确计算和报告效应量及其置信区间，本文首先介绍效应量的置信区间及其优势，然后以两种常用的效应量(Cohen’s

d 及 Eta squared，η^2）为例，介绍其置信区间的原理及如何在开源软件（如 R 和 JASP）中实现。但值得注意的是，本文提及的效应量并不仅限于 Cohen`s *d* 等标准化的效应量指标。根据 Cumming（2014）的定义，效应量指的是研究者感兴趣的任何效应的量，效应量既可以是标准化的，也可以是未标准化的、带有原始单位的。另外，并非标准化的效应量就一定优于未标准化的效应量，研究者应根据实际情况，选择能够合理反映数据信息的效应量，有时未标准化的效应量是更具解释力的。

2 报告效应量及其置信区间的优势

与 NHST 中的 *p* 值相比，报告效应量及其置信区间为结果提供了更详细、更多元的信息。具体而言，报告效应量及其置信区间有如下优势。

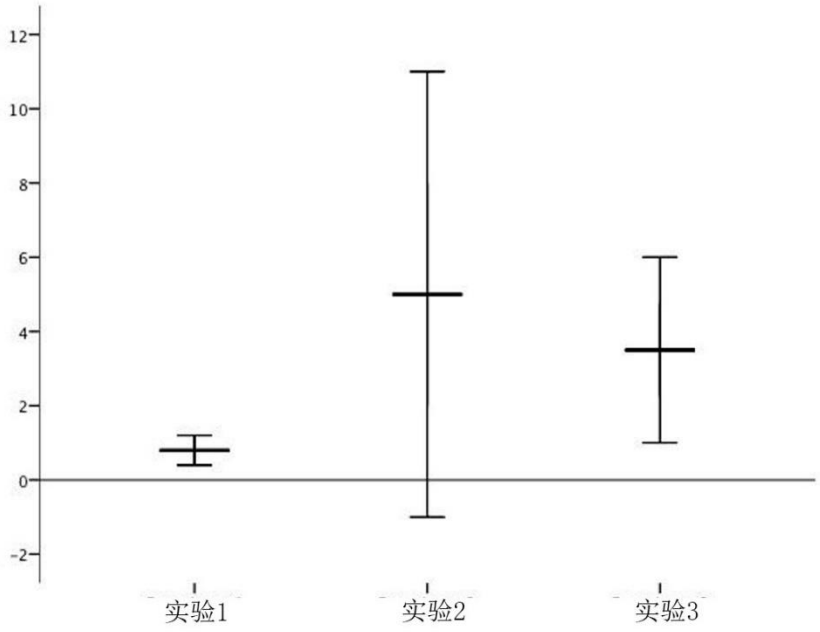

图 2 三组数据的效应量及其置信区间

第一，能够比较不同实验之间的误差大小。假如研究者进行了三个实验，其效应量及置信区间分别如图 2 所示。如果根据传统的 NHST 方法，研究者能够得出的结论为：在实验 1 和实验 3 中，$p < 0.05$，即两组均值均与 0 有显著差异；而实验 2 的 *p* 值大于 0.05，即其均值与 0 没有显著差异。在这种情况下，研究者根据实验 1 和实验 3 得出的结论几乎是相同的。至于两组均值的差异到底有多大？数据的抽样误差如何？三组数据哪一组能为假设提供最可靠的证据？*p* 值无法给出答案。

在传统的报告规范中，研究者通常利用未经标准化的点估计指标（例如：均值）及标准误来弥补上述不足，同样的报告效应量（此处为均值差）及其置信区间也能够达到相同目的。根据图 2 可知，实验 1 与实验 3 虽然均显著，但是相对而言，实验 1 的效应量较小、变异也较小，实验 3 则效应量较大，变异也较大。由于对效应量及其置信区间的分析，研究者对实验 1 和实验 3 的结论就会有所区别。

第二，效应量及其置信区间能帮助研究者得出正确的结论。在仅参考效应量及置信区间的情况下，大部分研究者能够在比较不同研究的结果时得出符合逻辑的结论；但仅凭 NHST 和效应量时，能够得出正确的结论的人数减少(Coulson, Healey, Fidler, & Cumming, 2010; Lyu et al., 2018)。相比 NHST 的二分思想，报告效应量及其置信区间将研究者引向一种“估计”、“定量”的取向(Cumming & Fidler, 2009)。在这种思维取向下，研究者也更倾向于提出量化的问题。仍以图 2 为例，实验 2 的结果虽然不显著，但是从效应量及其置信区间上来看，该实验的趋势与实验 1 和实验 3 是相同的。这也使得研究者对研究产生进一步深化的思考。例如，是否是实验 2 中数据的“噪音”过大导致了不显著的结果？

第三，可以展现关于研究的更丰富的信息。在图 2 中，实验 1 的效应量其实很小，换言之实验 1 中的两组实际差异可能不大。但是也许由于实验抽样误差小、样本量较大，实验 1 的置信区间很窄，研究者可以在很高的置信水平上得到差异显著的结论。这就是统计显著性与实际显著性不相称的实例。与之相反，对于实验 2，虽然其置信区间包含 0，但其效应量的点估计值却是最高的，由此可见在实验 2 数据的“噪音”过大，导致了其数据变异过大、置信区间过宽。实验 3 的结果则较为理想，其效应量及其置信区间都在较为合理的水平。

最后，由于效应量具有非样本依赖性(卢谢峰等, 2011)，相比依赖样本的 p 值，它更适用于跨实验的综合分析及元分析研究中。从频率主义统计的角度来讲，任何一个单独的研究可以看作是进行一次独立的抽样并对总体的参数进行一次估计，因此单个的研究很可能是片面的，但通过多个研究的数据积累，研究者可以进行通过元分析(meta-analysis)对总体进行更加精确地估计。元分析不仅能扩大样本量，提高统计检验力，还可以缩小置信区间的范围，使得对总体效应量的估计更加精确(Cumming, 2012)。相比 p 值，效应量及其置信区间的研究更便于进行元分析统计，且定量报告效应量及其置信区间的过程本身也蕴含了元分析思想。

正是由于效应量与置信区间的优势，其得到了研究者的广泛推荐。在美国心理学会（American Psychological Association, APA）出版手册（第六版）中，推荐了报告效应量及其置信区间。而在 2018 年《美国心理学家》（American Psychologist）所介绍的期刊报告标准中，也推荐报告效应量及其置信区间(Appelbaum et al., 2018)。

总之，在当前的研究中，虽然报告效应量及其置信区间得到了广泛的支持，但是效应量的置信区间却应用较少(Fritz et al., 2012)。一个主要的原因可能在于研究者都对效应量的置信区间知之不多，而且缺乏相应的工具进行实现（例如心理学常用的统计软件 SPSS 并没有常用效应量指标的置信区间输出）。为了解决这个问题，接下来，本文将以 Cohen’s d 和 Eta squared（η^2）为例，介绍其置信区间的原理与计算公式，并展示如何使用开源的软件来实现置信区间的计算。

3 标准化的差异量（Cohen’s d）

Cohen 最早对 d 的定义是以总体的标准差为标准化单位，然而在实际研究中总体的标准差常常是未知的，因此更常见的做法是使用样本的标准差作为标准化单位（后文以样本标准差 s 为单位进行描述）。Cohen’s d 的原理即为样本的均值和虚无假设（H_0）的均值差异除以

标准差的比值：

$$\text{Cohen's d} = {(\bar{X} - \mu)}/{s} \tag{3.1}$$

其中，*s* 表示样本的标准差，μ表示我们希望用来测量 *d* 的参考值。Cohen's *d* 就可以简单理解为样本均值$\bar{X}$与参考值μ之间相差几个标准差 *s*。不过，对比不同的研究目的，关于 Cohen's *d* 的计算公式有多种形式，具体可以参考 Cumming (2014)，Hedges (1981)和 Lakens (2013)。

3.1 Cohen's *d* 置信区间的原理

要理解 Cohen's *d* 的置信区间，首先需要理解 *t* 值在虚无假设（null hypothesis, H_0）为真（即没有效应）和备择假设（alternative hypothesis, H_1）为真这两种情况下的分布。假设从一个正态分布（$N(\mu, \delta)$）中随机抽取无数个样本量为 *N* 的样本。对于其中的一个样本，其均数为 *M*，标准差为 *s*。如果想检验这个样本是否属于标准正态分布的总体，在 NHST 的框架下，我们可以基于虚无假设H_0：$\mu = \mu_0$进行单样本 *t* 检验，可以通过如下公式计算 *t* 值：

$$t = \frac{\mathrm{M} - \mu_0}{s/\sqrt{N}} \tag{3.2}$$

在虚无假设为真的情况下，假如我们无数次进行抽取样本量为 N 的样本并进行 *t* 检验，那么这些 *t* 值会形成一个自由度$df = (N - 1)$的 *t* 分布。在这种情况下，*t* 分布是以 0 为中心，两边对称的分布。此时，我们也可以将 *t* 检验的统计量看作是 *M* 与μ之间以$s/\sqrt{N}$（标准误）为单位的距离。对于每一个样本，我们都可以使用 *t* 分布表计算 *p* 值，并进行假设检验。

但是，如果虚无假设（H_0）不为真，那么备择假设（H_1）即为真，即$\mu = \mu_1 (\mu_1 \neq \mu_0)$。在这种情况下，我们实际上是从均值为$\mu_1$的总体中进行抽样，那么无数次抽取样本量为 *N* 的样本而计算出来的均值 *M* 就会更加接近μ_1而非μ_0。如果仍用上面的公式进行 *t* 检验，那么无数次计算到的 *t* 值不再是以 0 为中心两侧对称的 *t* 分布，而是中心不在零点的偏态的非中心 *t* 分布。对于这样一个非中心 *t* 分布，其参数除了自由度（*df*）外，还包括一个非中心参数 Δ（读为：delta），Δ 可以看作是是μ_0和μ_1之间以标准误为单位的距离。在其他条件相同的情况下，Δ 值越大，说明这个非中心 t 分布的中心越偏离 0（如图 3 所示，其中非中心参数 ncp 表示 R 软件中 Δ 的取值）。

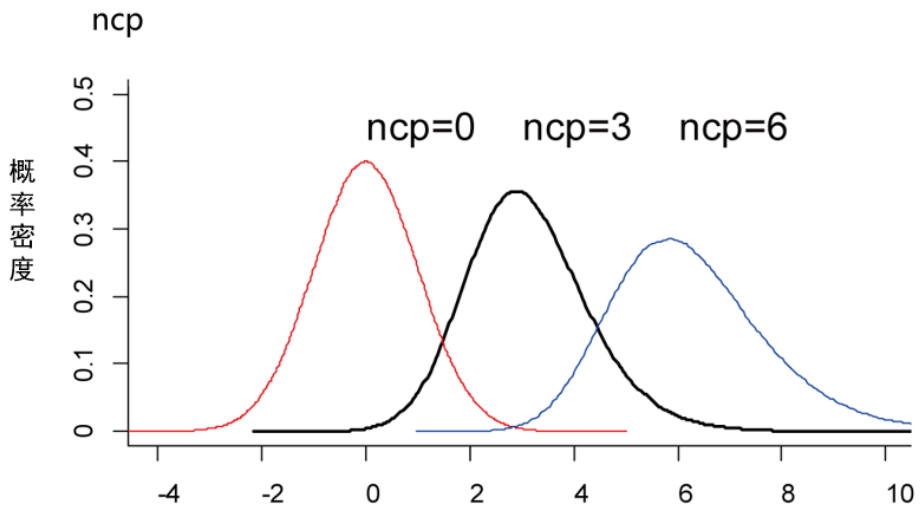

图 3 不同的非中心参数 Δ(ncp)对应的非中心分布 *t*

将公式(3.1)和公式(3.2)结合，可以得出

$$\text{Cohen's } d = {}^{t}/_{\sqrt{N}} \tag{3.3}$$

公式(3.1)说明 *d* 表示 *M* 与μ之间以 *s*（即标准差）为单位的距离；公式(3.2)说明 *t* 表示 *M* 与μ之间以$s/\sqrt{N}$（即标准误）为单位的距离。公式(3.3)则表明，Cohen's *d* 与 *t* 值有一一对应关系。因此，Cohen's *d* 的抽样分布也是非中心 *t* 分布，在计算 Cohen's *d* 的置信区间时需要用到非中心 *t* 分布。

由于 *t* 值在备择假设（H_1）为真时为非中心 *t* 分布，这种情况下 *d* 也是一个非中心 *t* 分布。也就是说 *d* 的置信区间是一个非对称的区间，上下限到中心的距离不一致，所以我们需要用迭代估计（iterative approximations）的方法来构建 *d* 的置信区间。我们可以结合下图来详细说明。

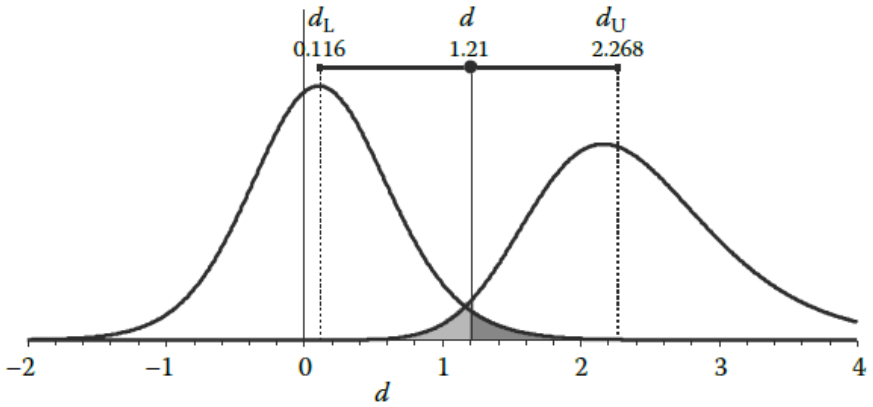

图 4 总体效应量 d 的可能值分布(引自 Cumming (2012), 第 11 章)

假如有一个总体效应为 Cohen's *d* = 1.21，需要构建其 95%的置信区间（如图 4 所示）。

也就是说，如果无数次构建这样的区间，约有 95%的区间包含 1.21。那么，以区间的下限 d_L 为中心时，d 的抽样分布拒绝 d_L 而选择真值的概率为 2.5%（深灰色部分）；同时，对于以置信区间上限 d_U 为中心时，d 的抽样分布拒绝 d_U 而选择真值的概率同样为 2.5%（浅灰色区域）。这就意味着，区间的上限和下限为中心的分布包含真值的可能性之和正好为 5%；而将区间下限或者上限向中心移动时，包含真值的可能性变大。同理，如果需要估计 99%置信区间的范围，相比于 95%的置信区间，区间的上限和下限会更远离中心，区间的上限和下限为中心的分布包含真值的可能性之和为 1%，那么深灰色部分和浅灰色部分应该是 0.005。

Exploratory Software for Confidence Intervals（ESCI）是由 Geoff Cumming 设计开发的一系列 Excel 文件，可以仅仅依托我们常用的 Microsoft Excel 软件完成复杂的统计计算，这其中包括效应量 Cohen's d 及其置信区间(Cumming, 2001)。使用 ESCI 可以更加直观地理解区间上限与下限与 d 值的关系。在 ESCI 中，将以区间下限 d_L 为中心的分布往左移动，d_L 就会变小，该分布右侧超过真值的区域也会变小；这意味着真值所对应的 p 值也会变小，那么能够拒绝 d_L 选择真值的概率就会变小。同样的，如果将以区间下限 d_L 为中心的分布往右移动，那么 d_L 值就会变大，该分布右侧超过真值的区域就会变大，那么能够拒绝 d_L 选择真值的概率就会变大。为了能得到一个准确的 95%的置信区间，我们需要移动以 d_L 为中心的分布使得它右侧超过真值的区域为 0.025，同时移动以 d_U 为中心的分布，使得它左侧超过真值的区域也为 0.025。这样得到的 d_L 和 d_U 就是我们需要的置信区间的上下限。

因为这两个曲线都是非中心 t 分布，所以我们可以改变 d 值来调整曲线向左右滑动。这种不断地调整以达到我们需要的区间的方法，即为迭代估计。简单来说就是在保持自由度不变的情况下，通过代入不同的非中心参数Δ（在一些研究中也会写作δ）进行相应的计算，并进行下一步的调整。在计算置信区间时，不断地调整Δ，从而不断调整非中心 t 分布，使得我们得到的在曲线上的临界值正好在 0.025 和 0.975 的双尾范围之间，这样我们就得到了 Cohen's d 的置信区间。那么，我们应该如何确定分别以置信区间上限和下限为中心的分布的非中心参数呢？

对于单样本的研究，非中心参数 Δ 的计算公式为

$$\Delta = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{N}}$$

结合公式(3.1)，我们就可以得到

$$\Delta = d\sqrt{N} \tag{3.4}$$

ESCI 使用公式(3.4)将 Cohen's d 和非中心参数 Δ 进行转换，而非中心参数 Δ 可以用来计算非中心的 t 分布。因此，我们可以得到 Cohen's d 的置信区间为：

$$d_L = \frac{\Delta_{.975}}{\sqrt{N}}$$

$$d_U = \frac{\Delta_{.025}}{\sqrt{N}}$$

相似的，对于双样本的研究，非中心参数 Δ 的计算公式为：

$$\Delta = \frac{\mu_2 - \mu_1}{\sigma\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}} \tag{3.5}$$

此时我们也可以依次计算双样本研究中的效应量和置信区间了：

$$\Delta = \frac{d}{\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}} \tag{3.6}$$

$$d_L = \Delta_{.975}\sqrt{\frac{1}{N_1}+\frac{1}{N_2}} \tag{3.7}$$

$$d_U = \Delta_{.025}\sqrt{\frac{1}{N_1}+\frac{1}{N_2}} \tag{3.8}$$

关于 Cohen's *d* 置信区间的原理，具体细节可参考 Cumming (2012) 第 11 章的内容。

3.2 实例与软件分析

在研究实践中，研究者不需要自己进行迭代来估计 Cohen's *d* 的置信区间。目前，R 语言(R Core Team, 2018)中有不少成熟的工具包可以用于计算 Cohen's *d* 的置信区间。而 JASP 是基于 R 所开发的用户界面友好的软件可以进行传统的统计分析和贝叶斯因子分析(Wagenmakers et al., 2015; 胡传鹏等, 2018)，也可以实现 Cohen's *d* 的置信区间的计算。（关于 SPSS 中计算 Cohen's *d* 置信区间的插件，见：http://dl.dropbox.com/u/1857674/CIstuff/CI.html；基于 Microsoft Excel 所开发的 ESCI 计算 Cohen's *d* 置信区间，见：https://thenewstatistics.com/itns/esci。）

我们将使用 JASP 示例数据"Kitchen Rolls" (具体数据，见：https://osf.io/q9387/) 进行说明。Topolinski 和 Sparenberg(2012)发现，转动纸卷的方向能够改变个体在人格量表上开放性的得分，Wagenmakers 等(2015)对此实验进行重复实验，这里使用的数据即为 Wagenmakers 等(2015)的重复实验数据。该示例数据包含两组被试在人格量表中关于开放性的得分，其中一组被试在填写问卷时顺时针旋转桌面上的纸卷，而另一组则逆时针旋转。数据分析中，NEO PI-R 的平均得分作为因变量，被试的分组（顺时针或逆时针）为自变量，采用独立样本 *t* 检验进行数据分析。

3.2.1 使用 JASP 计算 Cohen's *d* 的置信区间

将样例数据使用 JASP 打开后，选择 T-Tests → Independent Samples T-Test, 得到如下界面。根据要求将需要统计的变量导入对应变量框中（与 SPSS 类似），在下方界面点选需要进行的统计操作，其中在 Additional Statistics 下可以勾选 Effect Size 和 Confidence interval 的选项，根据公式(3.5)-(3.8)计算结果即为效应量 Cohen's *d* 及其置信区间。

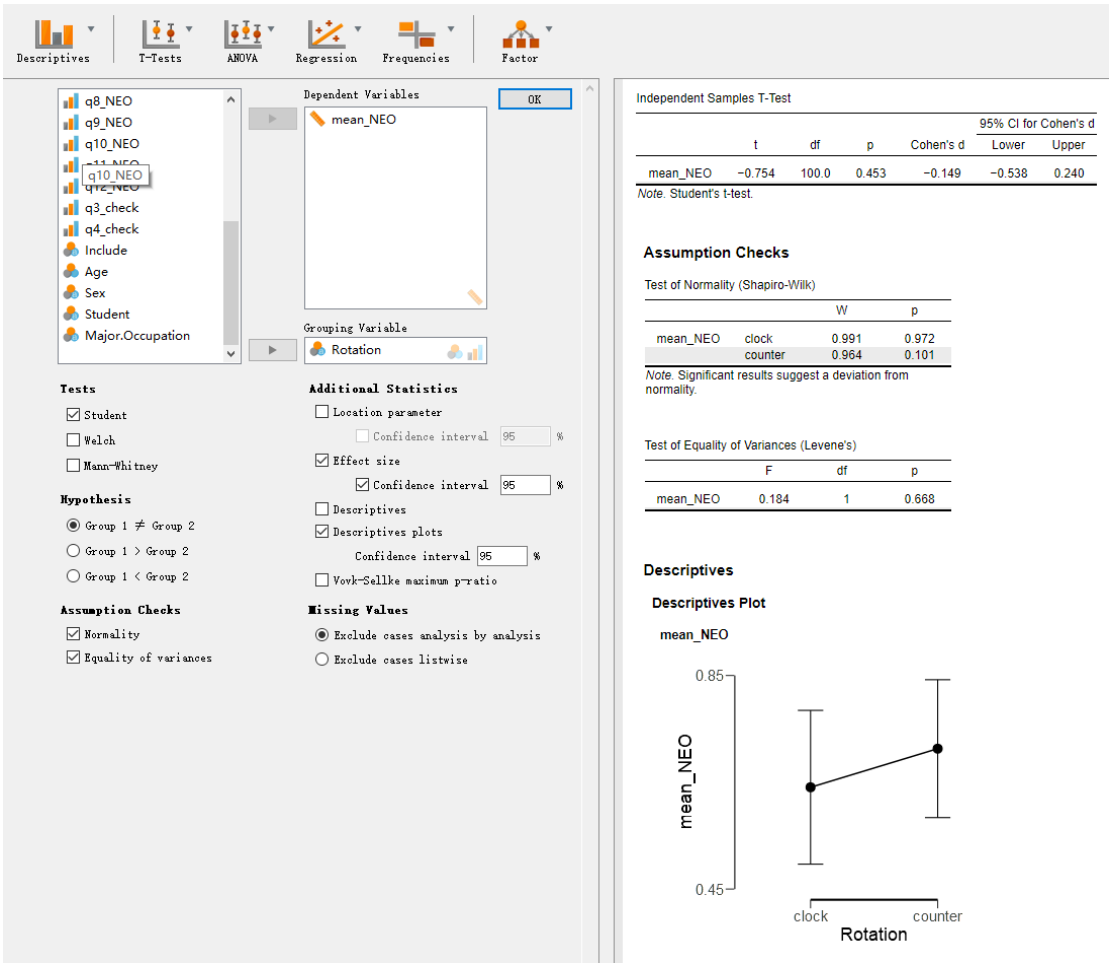

图 5 JASP 独立样本 *t* 检验操作（左侧）及结果界面（右侧）

结果显示因变量满足正态分布和方差齐性假设，因此选择 Student *t* test 进行分析。结果显示两组的 NEO PI-R 的平均得分没有显著差异（$t(100) = 0.754, p = 0.453$），Cohen's $d = 0.149$, 95% CIs [-0.240, 0.538]。

3.2.2 使用 R 计算 Cohen's *d* 的置信区间

R 语言中有多个工具包可以完成独立样本 *t* 检验，如 car 和 MBESS。假如我们使用 car 工具包上的 t.test 函数，得到两组被试在 NEO PI-R 的平均得分没有显著差异，$t(100) = 0.754$, $p = 0.453$（当然，也可以使用 JASP 或者 SPSS 得到 *t* 值与 *p* 值）。在得到 *t* 值之后，则可通过使用如下命令来计算 Cohen's *d* 的置信区间，R 代码如下：

```
library("MBESS") # 打开 MBESS 工具包
#定义相关参数并计算 Cohen's d 的 95%置信区间
MBESS:: ci.smd(ncp = 0.75361, n.1 = 48, n.2 = 54, conf.level = 0.95)
```

其中 ncp（非中心参数）是 *t* 值，n.1 和 n.2 代表两组的样本量，MBESS 采用公式(3.5)-(3.8)通过运行程序可以获得结果。

3.3 结果报告与解释

如上所示，使用两种不同的软件对于顺时针旋转组的被试与逆时针旋转组的被试的人格量表得分差异进行估计，并且得到了 95%的置信区间。输出的结果都表明，两组被试的 NEO PI-R 的平均得分没有显著差异，对于效应量及其 95%的置信区间的估计也是相同的——效应量 *d* 为 0.149，其 95%置信区间为[-0.240, 0.538]。基于这些结果，我们可以得到的结论：目前的数据无法拒绝零假设，即无法推断出被试进行顺时针旋转或者逆时针旋转对于 NEO PI-R 的得分存在显著影响的。（注意，这里 $p > 0.05$ 及 Cohen's *d* 的置信区间包含 0 均无法得到*零假设为真*的结论，即无法使用 *p* 值来支持两组没有差异的结论，因为 *p* 值的计算是以零假设为真作为前提条件的。要为*零假设为真*这个结论提供证据，需要借助其他的统计手段。）

4 方差分析中效应量及其置信区间

心理学研究中另一个最为常见的效应量指标是方差分析（analysis of variance, ANOVA）中的 Eta-squared（η^2） (Fritz et al., 2012)，其最早由 Pearson (1905)提出，可以理解为单个或者多个因素（交互作用）引起的变异在总变异中所占的比例(Cohen & Cohen, 2010)。η^2的计算公式如下：

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{4.1}$$

非常值得注意的是，SPSS 输出的效应量指标 η_p^2在心理学研究中应用广泛，但是意义与 η^2不完全相同并且容易引起误解。例如有研究指出很多研究者很容易混淆 η^2和 η_p^2，这种混淆可能会造成一些比较严重的后果，如在元分析（meta-analysis）中如果错误的使用 η_p^2代替 η^2，会使得元分析结果出现严重的偏差(Levine & Hullett, 2002)。此外误用 η^2和 η_p^2对理论的建构也十分不利(Pierce, Block, & Aguinis, 2004)。因此报告 η_p^2的时候一定要注明报告的是哪个指标（对论文中 η^2与 η_p^2不明确情况下，可对各个影响因素的效应量相加，一般结果等于 1 的情况下是 η^2 ，如果结果大于 1，则是 η_p^2）。另外在样本量比较小的时候（自变量和样本的比值小于 1:10），ω^2则成为研究者更为推荐报告的效应量指标(卢谢峰等, 2011)。当然与 ω^2类似的效应量统计指标还有 ε^2，详见(Maxwell & Delaney, 2004)。下面结合公式 4.1 主要对 η^2置信区间计算进行说明。

4.1 η^2置信区间计算的原理

要理解 η^2的置信区间，同样需要理解与其相关参数有关的非中心性分布。在这里，η^2置信区间的建构需要方差分析中 *F* 值的分布以及方差分析中另一个效应量指标 Cohen's *f*。以最简单的单因素被试间设计方差分析为例，其总体变异可以被分解成为组间变异和组内变异：

$$SS_{total} = SS_{between} + SS_{error}$$

即：

$$\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\bar{X})^2=\sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2+\sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)^2$$

其中 X 表示观测值，j 表示分组水平（共有 k 组），n 表示组内被试数量（每组内均有 n 个被试）。此时，F 值计算公式如下：

$$F(df_1,df_2)=\frac{SS_{between}/df_1}{SS_{error}/df_2} \tag{4.2}$$

其中$df_1=k-1, df_2=nk-df_1-1$。此时组间处理的效应量为：

$$\eta^2=\frac{SS_{between}}{SS_{total}}=\frac{SS_{between}}{SS_{between}+SS_{error}}=\frac{\sum_{j=1}^{k}(\bar{x}_j-\bar{x})^2}{\sum_{j=1}^{k}\sum_{i=1}^{n}(x_{ij}-\bar{x})^2} \tag{4.3}$$

组间效应的另一种效应量指标——Cohen's f 则可以通过如下公式计算：

$$f=\sqrt{\frac{SS_{between}}{SS_{error}}}=\sqrt{\eta^2/(1-\eta^2)} \tag{4.4}$$

此时的 F 分布和χ^2 分布存在非常紧密的关联。根据χ^2 分布的定义可知，χ^2 分布是从标准正态分布中独立抽取出样本的平方和的分布。也就是说，假设有 p 个从标准正态分布（N(0,1)）中抽取出来的随机变量集合$\{X_i, i=1,\dots, k\}$，则有：

$$\sum_{i=1}^{k}(X_i^2-\mu)/\sigma^2=\sum_{i=1}^{k}(X_i)^2$$

这是一个自由度为 k-1 的χ^2 分布，且这个χ^2 分布是中心性的(注意，这里的中心性并非指的是该分布是中心对称，而是说其是从中心对称的分布中抽出来的数据的平方和的分布)。对照之前方差分析中 F 值的计算公式，如果将分子和分母同时除以$\sigma^2_{between}$（处理引起的变异）和σ^2_{error}（误差引起的变异）(在 ANOVA 的 H_0 为真的情况下，假设处理变异同误差引起的变异相同即$\sigma^2_{between}=\sigma^2_{error}$，所以在公式中相互抵消了)，则 F 值（$F(df_1, df_2)$，以下简写为 F）的分子和分母分别对应一个χ^2 分布。

$$F(df_1,df_2)=\frac{(SS_{between}/df_1)/\sigma^2_{\text{between}}}{(SS_{error}/df_2)/\sigma^2_{\text{error}}}$$

在 ANOVA 中，由虚无假设为组间均数相等，实验误差服从正态分布 $N(0,\sigma_{error})$可知，此时的分子分母对应的χ^2 分布是中心性。在此这情况下，F 分布也呈中心性。

当虚无假设为假的时候，组间均数不相等，分子对应的χ^2分布呈非中心性，分母作为实验误差对应的分布还是中心性的χ^2分布。此时的 F 分布也变成了非中心性的，可以表示为$F(df_1, df_2, \delta)$。实际上，中心分布是非中心分布的特殊情况。非中心参数 ncp 决定了分布的具体形态，例如中心 $F(2, 52, ncp = 0)$分布（黑色）和非中心 $F(2, 52, ncp = 1)$分布（红色），如下图所示。

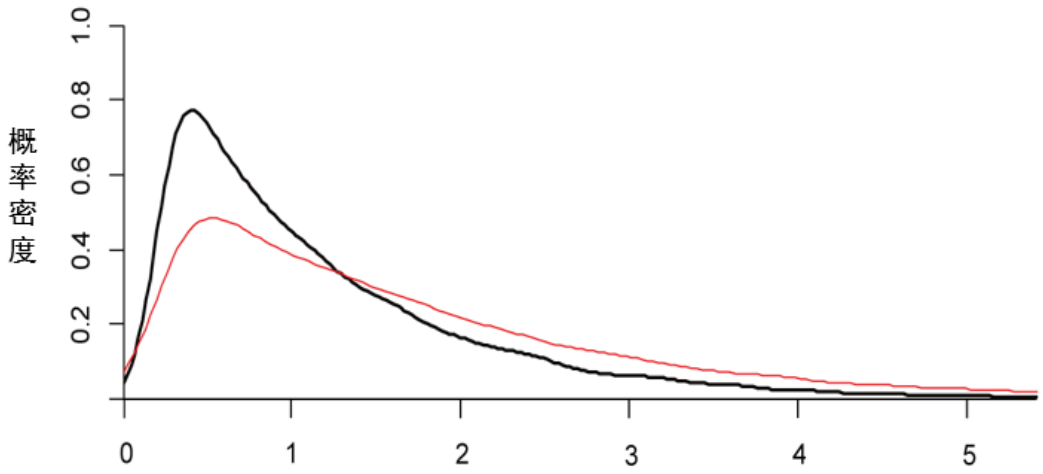

图 6 中心 F 分布和非中心 F 分布

计算效应量的前提就是承认 H_0 为假（组间均数不相等），其对应的 F 分布是非中心分布。如果计算 η^2的置信区间是基于非中心 F 分布，则其区间估计的上下限过程中，存在与 Cohen's d 置信区间估计过程中同样的问题：在置信区间的上限与下限位置的 F 分布的非中心参数不相同。因此，对于 η^2的置信区间的估计，同样需要使用反演原理（inversion confidence interval principle）(Steiger & Fouladi, 1996)。

我们通过三个阶段得到置信区间：统计检验→ 非中心参数→ 效应量统计。首先我们需要建立统计检验值（方差分析下的 F 值）和非中心参数以及效应量 η^2之间的关系。由公式 4.3 可得$f^2 = \frac{SS_{between}}{SS_{error}}$，因此，可以推出

$$F(df_1, df_2) = f^2(df_2/df_1) \tag{4.5}$$

当虚无假设为假时，$F(df_1,df_2)$的非中心参数的估计值 δ (非中心参数的符号表述方式可能会有不同 ，常用的符号包括δ、λ)的计算公式如下(Smithson, 2001)：

$$\delta = f^2(df_1 + df_2 + 1) \tag{4.6}$$

结合公式（4.5），我们得到非中心参数的估计：

$$\delta = [F * (df_1/df_2) * (df_1 + df_2 + 1)] \tag{4.7}$$

至此我们建立起了统计值 F 和非中心参数之间的关系。再综合公式（4.2），（4.3）和（4.7），可以推断出 η^2 与 f^2 和非中心参数 δ 的关系如下：

$$\eta^2 = f^2/(1+f^2) = \delta/(\delta + df_1 + df_2 + 1) \tag{4.8}$$

至此，我们得到了 η^2 与 F 值、F 分布的非中心参数之间的关系。接下来，我们就可以使用置信区间反演原理来计算 η^2 的置信区间。假设给定我们一个样本 $F(5,194)$，我们需要构建一个 100(1-α)%（α=0.05）的双侧的置信区间（如图 7 所示）。

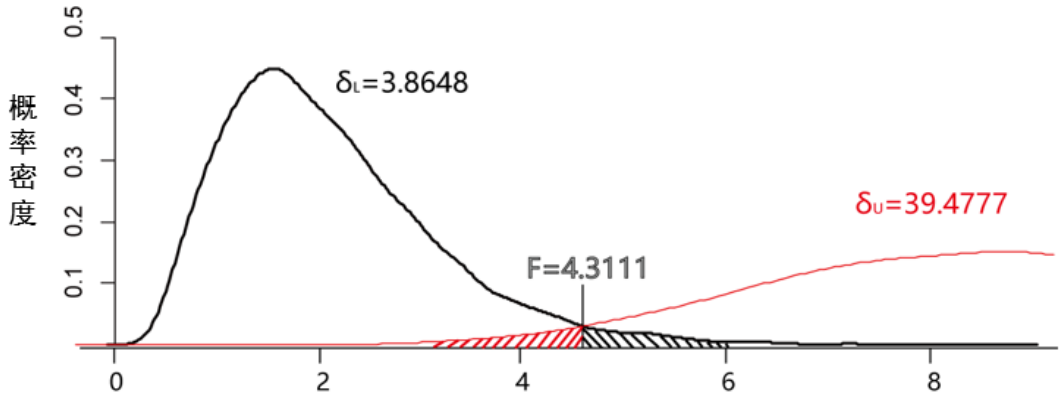

图 7 构建 η^2 置信区间的示例图 (Smithson, 2001)

下限 δ_L 对应 $F(5,194)$ 右侧的α/2 处，上限 δ_U 对应 $F(5,194)$ 左侧的α/2 处。在得到与上下限对应的非中心参数 δ 后，我们可以将其转换为 η^2 的置信区间，转换公式如下：

$$\eta_L{}^2 = \delta_L/(\delta_L + df_1 + df_2 + 1) \tag{4.9}$$

$$\eta_U{}^2 = \delta_U/(\delta_U + df_1 + df_2 + 1) \tag{4.10}$$

这样我们就完成了对 η^2 的置信区间的估计。

值得注意的是，对 ANOVA 效应量置信区间的计算，通常报告 90%的置信区间即可。原因在于均值之间的差异可以是正值也可以是负值，但是由于 η^2 或 R^2 是平方值，所以只有正值。计算 95%的置信区间时，可能会得到包含 0 的置信区间，但此时 p 值可能小于.05，此时置信区间的结果与 p 值出现了矛盾（见 Karl Wuensch 的解释：http://core.ecu.edu/psyc/wuenschk/spss/spss-programs.htm）。而且 Steiger（2004）指出均值比较的 95%置信区间和 90%置信区间得到的检验效力是一样的，并且 η^2 不可能小于 0，所以与 0 不存在显著差异的置信区间（通常情况下不包含 0）的下限至少要从 0 开始(Steiger, 2004)。

4.2 η^2 及其置信区间在 R 上的实现

同样，我们将采用由 JASP 提供的样例数据来演示如何使用 R 计算 η^2 的 90%CI。该数

据名为 Tooth Growth 和 Bugs，分别用来展示被试间设计和被试内设计方差分析中 η^2及其 CI 的实现（SPSS 上如何实现，见：http://core.ecu.edu/psyc/wuenschk/spss/spss-programs.htm）。

4.2.1 被试间设计 η^2及其置信区间在 R 上的实现

Tooth Growth 数据来自两因素完全随机设计，60 只豚鼠被随机分配到 6 种处理条件下，用以研究不同类型的营养品（维生素 c 即 VC 和橙汁 OJ）在不同抗坏血酸剂量条件下（0.5mg、1mg 和 2mg）对豚鼠牙齿生长的影响，因变量选取的是豚鼠牙齿的长度。

首先使用统计软件获得计算置信区间所需的统计值。这里你可以使用 R 中自带的函数 aov 或者一些带统计功能的工具包（如 ez、car 等等），这里需要注意的是用 R 进行方差分析时，不同的工具包或者函数使用的平方和类型会有所不同，例如 aov 函数进行计算的时候默认使用的是 Type I SS（sun of square），ezANOVA 默认使用的是 Type II SS（可以在 R 中使用 type 对平方和类型进行调整，详见 https://cran.r-project.org/web/packages/ez/ez.pdf），而 SPSS 在进行方差分析计算的时候默认的是 Type III SS（可以在 SPSS 中模型选项进行调整）。当数据不同组间的被试量相同时，不同类型平方和计算结果出现的差异不大，但是当数据不平衡的时候，则要谨慎考虑平方和类型，因为不同的平方和类型会带来不同的统计结果，感兴趣的读者可以参考(Langsrud, 2003)。当然更为便捷的办法是应用 JASP 直接进行统计分析并获得相应的统计值。例如对于以上数据，可得 $F(2,54) = 92$，随后在 R 中下载并打开 MBESS 工具包，输入相关的统计值进行置信区间的计算，R 中的命令如下：

```
library("MBESS") # 打开 MBESS 工具包
ci.pvaf(F.value=92,df.1=2,df.2=54,N=60,conf.level=.90) # 输入 F 值、自由度计算对应的 90%置信区间
```

4.2.2 被试内设计 η^2及其置信区间在 R 上的实现

Bugs 数据来自两因素混合设计，用以研究不同性别（男、女）人群对于不同类型（不吓人不恶心、不吓人很恶心、很吓人不恶心和很吓人很恶心）虫子图片的敌意指数，并采用 10 点评分表明想要杀死或者驱赶虫子的程度(Ryan, Wilde, & Crist, 2013)。通过 JASP，我们可以得到 $F(2.64, 224.48)$，（注意被试内设计数据在违背球形假设的情况下使用校正后的自由度）。然后再 R 中使用如下命令得到置信区间：

```
# 打开 MBESS 工具包
library("MBESS")
# 输入 F 值及自由度
Lims<-conf.limits.ncf(F.value=20.14,conf.level=0.90,df.1=2.64,df.2 = 224.48)
# 计算 90%置信区间的下限
```

```
Lower.lim<-Lims$Lower.Limit/(Lims$Lower.Limit+df.1+df.2+1)
# 计算 90%置信区间的上限
Upper.lim<-Lims$Upper.Limit/(Lims$Upper.Limit+df.1+df.2+1)
```

4.3 结果报告与解释

对于 η^2及其置信区间的解释主要参照 η^2的定义，也就是实验效应引起的变异占总体变异的比例，因此 η^2的大小说明了在具体的实验研究中对于自变量操作的有效性。也就是说 η^2越大，相关变量之间的关系越紧密，当然这种关系的属性，即相关还是因果关系主要由实验设计的类型（如准实验设计和实验设计）决定。但是由于 η^2置信区间不可能小于 0，这也就决定了对于 η^2的解释不可能像前面提到的 Cohen's *d* 的置信区间一样，把包含 0 的置信区间作为我们拒绝或者接受零假设的依据。而且方差分析的应用作为一般线性模型下的特例，往往只是对涉及变量间关系检验的第一步。因此我们一般把 η^2及其置信区间作为评价实验变量操控有效性的指标，接下来具体的组间比较才是研究者关注的重点（例如主效应显著后的多重比较、交互作用显著后的简单效应分析），而在组间比较中可以再次使用如 *t* 检验下的 Cohen's *d* 作为评价组间差异可靠性的效应量指标。

5 总结

近年来心理学中的可重复危机已经对心理学界产生了深远的影响，而统计报告标准的变化，组成了期刊论文报告标准变化中非常重要的部分(刘宇等, 2018; Appelbaum et al., 2018; Levitt, Bamberg, Creswell, Frost, Josselson, & Suárez-Orozco, 2018)。Cohen's *d* 与 η^2作为基于估计统计中两个最常用的效应量指标，对于研究者来说具有重要意义(Fritz et al., 2012)。本文解释了这两个效应量置信区间的原理，并采用实例演示了如何在 R 与 JASP 中实现这两种置信区间（所有演示数据与代码，见：https://osf.io/4ameb/），可能对研究者具有一定的帮助。虽然本文未对另一个常见的效应量指标——相关系数的置信区间也进行说明及演示，但是其计算与实现在 JASP 与 R 中均相对成熟，读者可以参阅相关资料。更多关于置信区间的原理，可见(Smithson, 2003)。

值得注意的是，任何一个统计方法均有其优缺点(Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016)。对于心理科学而言，任何新的统计方法都不足以解决可重复危机(胡传鹏等, 2016；刘佳, 霍涌泉, 陈文博, 解诗薇, 王静, 2018)。对于研究者以及整个领域来说，最重要的是充分理解各个统计方法的前提及其不足，否则难以真正避免假阳性。本文所介绍的内容，可能可以帮助研究者达到新报告标准的要求，在结果中提供更丰富的信息。

参考文献

胡传鹏, 孔祥祯, Wagenmakers, E.-J., Ly, A., 彭凯平 (2018). 贝叶斯因子及其在JASP中的实现. *心理科学进展, 26*(6), 951–965.

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平 (2016). 心理学研究中的可重复性问题：从危机到契机. *心理科学进展, 24*(9), 1504–1518.

胡竹菁 (2010). 平均数差异显著性检验统计检验力和效果大小的估计原理与方法. *心理学探新, 30*(1), 68–73.

刘佳, 霍涌泉, 陈文博, 解诗薇, 王静 (2018). 心理学研究的可重复性“危机”: 一些积极应对策略. *心理学探新, 38*(1), 86–90.

刘宇, 陈树铨, 樊富珉, 邸新, 范会勇, 封春亮, . . . 胡传鹏 (2018). 心理研究的元分析报告标准：现状与建议. ChinaXiv. Retrieved from http://www.chinaxiv.org/abs/201809.00177

卢谢峰, 唐源鸿, 曾凡梅 (2011). 效应量:估计、报告和解释. *心理学探新, 31*(3), 260–264.

郑昊敏, 温忠麟, 吴艳 (2011). 心理学常用效应量的选用与分析. *心理科学进展, 19*(12), 1868–1878.

Publications, A. P. A., on Journal, C. B. W. G., & Standards, A. R. (2008). Reporting standards for research in psychology: Why do we need them? What might they be?. *The American Psychologist*, *63*(9), 839-851.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational & Psychological Measurement, 33*(1), 107–112.

Cohen, J., & Cohen, P. (2010). Applied multiple regression/correlation analysis for the behavioral sciences. *Journal of the Royal Statistical Society, 52*(4), 691–691.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Psychology, 1*: 26.

Cumming, G. (2001). *Project design and achieving educational change: from Statplay to ESCI*. Melbourne: Biomedical Multimedia Unit, The University of Melbourne,.

Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The New Statistics: Why and how. *Psychological Science, 25*(1), 7–29.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für*

Psychologie/Journal of Psychology, 217(1), 15–26.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., . . . Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*(3), 230–232.

Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, *1*(1), 60-69.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics, 6*(2), 107–128.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin Review, 21*(5), 1157–1164.

Kline, R. B. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. Washington, DC: American Psychological Association.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*: 863.

Langsrud, Ø. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. *Statistics & Computing, 13*(2), 163–167.

Levine, T. R., & Hullett, C. R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research, 28*(4), 612–625.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*,*73*(1), 26.

Lyu, Z., Peng, K., & Hu, C. P. (2018). P-value, Confidence Intervals and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology*, *9*: 868.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: a model comparison perspective*. New York: Routledge.

Pearson, K. (1905). *Mathematical contributions to the theory of evolution: XIV. On the general theory of skew correlations and nonlinear regression (Draper's Company Research Memoirs, Biometric Series II)*. London: Dulau.

Pedhazur, E. J., & Kerlinger, F. N. (1973). *Multiple regression in behavioral research: explanation and prediction*. New York: Holt, Rinehart and Winston.

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary Note on Reporting Eta-Squared Values from Multifactor ANOVA Designs. *Educational & Psychological Measurement, 64*(6), 916–924.

R Core Team. (2018). R: A language and environment for statistical computing. R foundation for statistical computing. Vienna , Austria. Retrieved from https://www.R-project.org/

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2016). Is there a free lunch in inference?. *Topics in Cognitive Science*, *8*(3), 520-547.

Ryan, R. S., Wilde, M., & Crist, S. (2013). Compared to a small, supervised lab experiment, a large, unsupervised web-based experiment on a previously unknown effect has benefits that outweigh its potential costs.*Computers in Human Behavior*, *29*(4), 1295-1301.

Smithson, M. J. (2003). *Confidence Intervals*. Thousand Oaks, CA: Sage.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182.

Steiger, J. H., & Fouladi, R. T. (2016). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 221-257). Mahwah, NJ, USA: Lawrence Erlbaum Assoc Inc.

Wagenmakers, E. J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., ... & Gronau, Q. F. (2015). Turning the hands of time again: a purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, *6*: 494.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35–57.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426–432.

Calculating Confidence Intervals of Cohen's d and Eta-squared: A Practical Primer

WANG Jun[1] SONG Qiongya[1] XU Yuepei[2] JIA Binbin[3] HU Chuan-Peng[4,5]

([1]*Department of Psychology, Sun Yat-Sen University, Guangzhou, 510006, China*)

([2]*College of Education, Shanghai Normal University, Shanghai, 200234, China*)

([3]*Shanghai University of Sport, Shanghai, 200438, China*)

([4] *Neuroimaging Center, Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Centre Mainz, 55131 Mainz, Germany*)

([5] *Deutsches Resilienz Zentrum (DRZ), University Medical Centre of the Johannes Gutenberg University, 55131 Mainz, Germany*)

Abstract

The recent replication crisis in psychology has motivated many researchers to reform the methods they used in research, reporting effect sizes (ES) and their confidence intervals (CIs) becomes a new standard in mainstream journals. However, a practical tutorial for calculating CIs is still lacking. In this primer, we introduced theoretical basis of CIs of the two most widely-used effect size, Cohen's *d* and η^2, in plain language. The CIs of both Cohen's *d* and η^2 are calculated under the condition that the alternative hypothesis (H_1) is true, and both rely on the estimation of non-centrality parameters of non-central distributions by using iterative approximations. More specifically, non-central *t*-distribution for Cohen's *d* and non-central *F*-distribution for η^2. Then, we illustrated how to calculate them in R and JASP with real data. This practical primer may help Chinese psychological researchers understand the CIs better and report CIs in their own research.

Key words: Effect size; Confidence interval; Cohen's *d*; Eta squared; R